Experiences and opportunities for one-sided communication in the ECMWF weather forecasting model

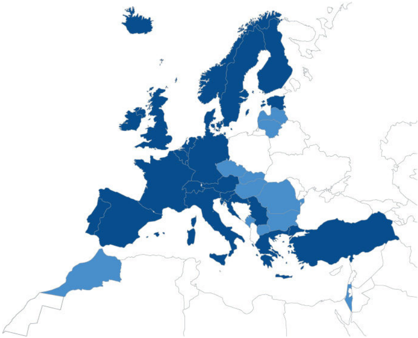
Ioan Hadade

ioan.hadade@ecmwf.int

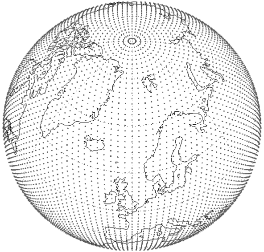
European Centre for Medium-Range Weather Forecasts

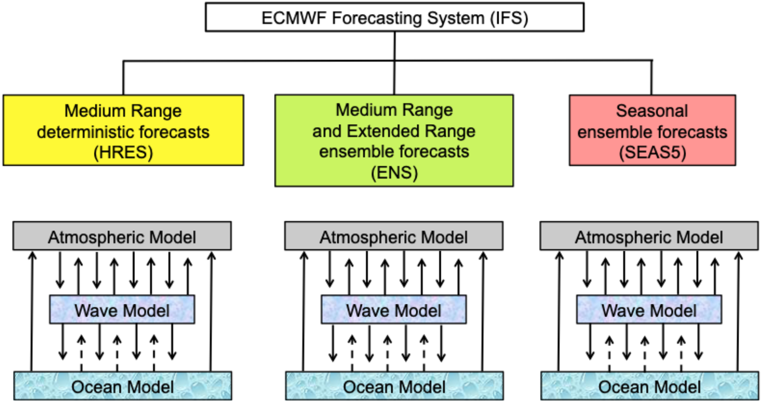
The European Centre for Medium-Range Weather Forecasts

- Independent intergovernmental organisation
- Established in 1975, today supported by **23 member** and **12 cooperating states**
- Headquarters in **Reading (UK)**, data center in **Bologna (IT)** and offices in **Bonn (DE)**
- Research institute and 24/7 operational service:
  - produce and disseminate global NWP products
  - operate meteorological data archive
  - implement Copernicus services CAMS and C3S
  - provide computing resources to member states



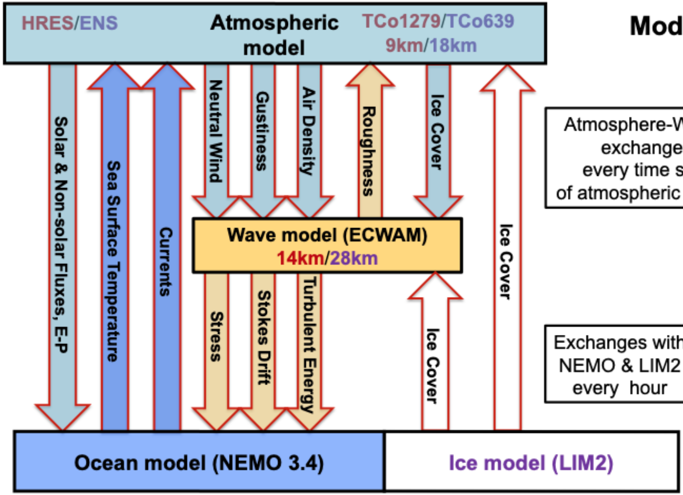
Configurations





Forecast period	10 days	15 days / 46 days	7 months / 13 months
Ensemble members	1	51	51 / 15
Frequency	2x daily	2x daily / 2x weekly	monthly / quarterly
Resolution	TCo1279L137 (9km)	TCo639L137 (18km) / TCo319L137 (36km)	TCo319L91 (36km)

Model components



Modeling infrastructure all models

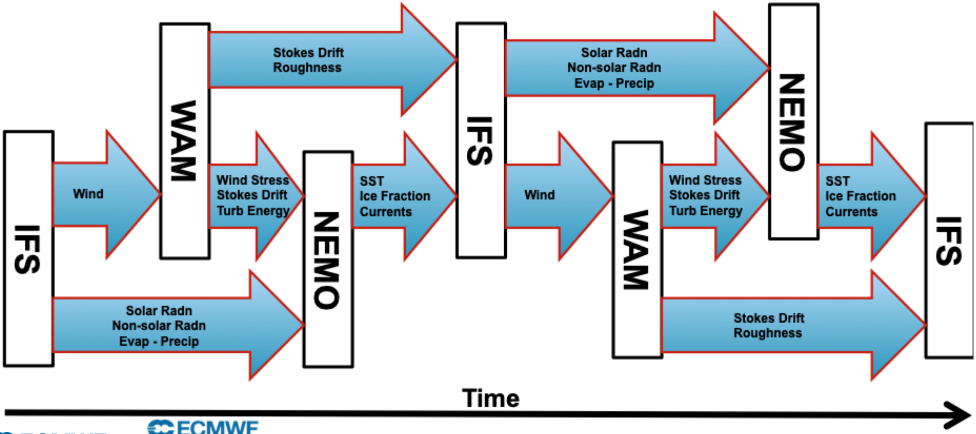
Atmosphere-Waves exchanges every time step of atmospheric model

Exchanges with NEMO & LIM2 every hour

Operational configuration Cycle 45R1 (from early 2018)

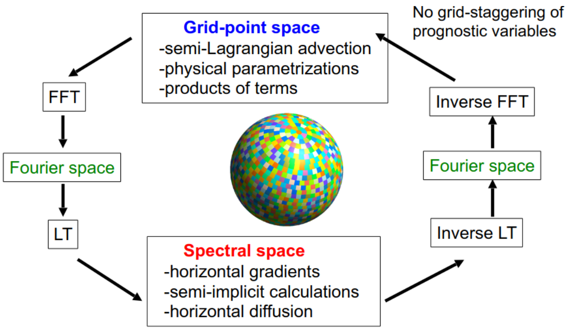
Time step view model components

Sequence of exchanges of physical quantities in IFS



Atmospheric model ("IFS")

Schematic of a time step:

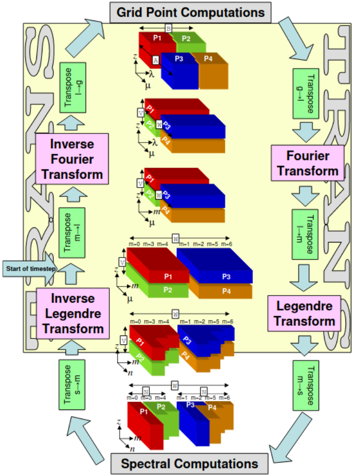


FFT: Fast Fourier Transform, LT: Legendre Transform

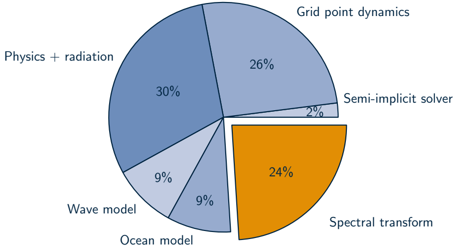
- Main component, ca. 80% of runtime
- F90, some F03/08, C++ libraries
- MPI and OpenMP

- **Dynamical core**
  - Hydrostatic model
  - Time stepping: semi-Lagrangian, semi-implicit
  - Horizontal discretization: Spectral transform method
  - Vertical discretization: Cubic spline finite elements
- **Physical parameterizations**
  - Tendencies from subgrid-scale processes

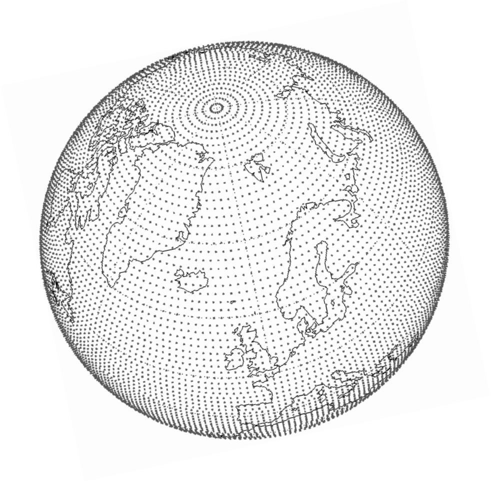
Spectral transforms



- Transformation from grid-point space to spectral space and back in **every time step**
- **Global communication** for every transpose

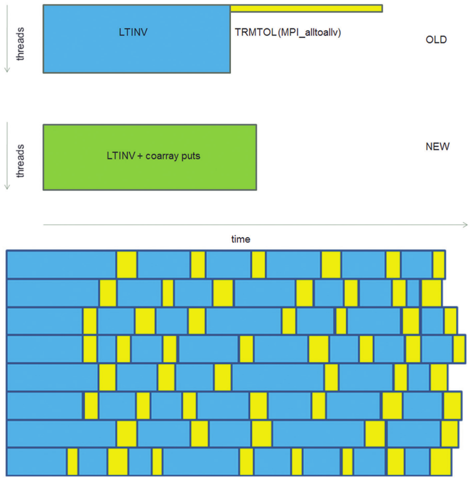


Domain decomposition



One-sided comms in spectral transforms

- Fortran2008 Coarrays and GASPI implementations
- Improve on the hybrid MPI/OpenMP funnelled approach by exposing parallelism at a finer granularity (within OpenMP parallel loops)
- Aim was to achieve improved overlap of computation and communication



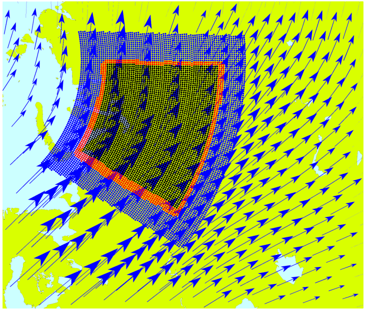
ORIGINAL

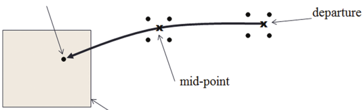
```
! COMPUTE
$OMP PARALLEL DO SCHEDULE(DYNAMIC,1) &
   & PRIVATE(JM,IM)
DO JM=1,D%NUMP
   IM = D%MYMS(JM)
   CALL LTINV(IM,JM,KF_OUT_LT,KF_UV, &
    & KF_SCALARS,KF_SCDERS,ILEI2,IDIM1,&
    & PSPVOR,PSPDIV,PSPSCALAR ,&
    & PSPSC3A,PSPSC3B,PSPSC2 , &
    & KFLDPTRUV,KFLDPTRSC,FSPGL_PROC)
ENDDO
!$OMP END PARALLEL DO
! COMMUNICATION
DO J=1,NPRTRW
   ILENS(J) = D%NLTSFTB(J)*IFIELD
   IOFFS(J) = D%NSTAGT0B(J)*IFIELD
   ILENR(J) = D%NLTSGTB(J)*IFIELD
   IOFFR(J) = D%NSTAGT0B(D%MSTABF(J))*IFIELD
ENDDO
CALL MPL_ALLTOALLV(PSENDBUF=FOUBUF_IN, &
 & KSENDCOUNTS=ILENS,&
 & PRECVBUF=FOUBUF,KRECVCOUNTS=ILENR,&
 & KSENDDISPL=IOFFS,KRECVDISPL=IOFFR,&
 & KCOMM=MPL_ALL_MS_COMM,CDSTRING='TRMTOL:')
```

NEW

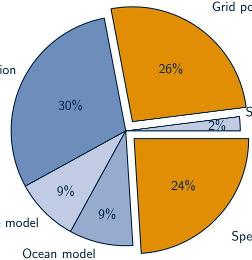
```
! COMPUTE
!$OMP PARALLEL DO SCHEDULE(DYNAMIC,1) &
  & PRIVATE(JM,IM,JW,IPE,ILEN,ILENS,IOFFS,IOFFR)
DO JM=1,D%NUMP
   IM = D%MYMS(JM)
   CALL LTINV(IM,JM,KF_OUT_LT,KF_UV, &
    & KF_SCALARS,KF_SCDERS,ILEI2,IDIM1,&
    & PSPVOR,PSPDIV,PSPSCALAR ,&
    & PSPSC3A,PSPSC3B,PSPSC2 , &
    & KFLDPTRUV,KFLDPTRSC,FSPGL_PROC)
! COMMUNICATION
  DO JW=1,NPRTRW
    CALL SET2PE(IPE,0,0,JW,MYSETV)
    ILEN = D%NLEN_M(JW,1,JM)*IFIELD
    IF( ILEN > 0 )THEN
      IOFFS = (D%NSTAGT0B(JW)+ &
            &  D%NOFF_M(JW,1,JM))*IFIELD
      IOFFR = (D%NSTAGT0BW(JW,MYSETW)+ &
            & D%NOFF_M(JW,1,JM))*IFIELD
      FOUBUF_C(IOFFR+1:IOFFR+ILEN)[IPE]= &
            & FOUBUF_IN(IOFFS+1:IOFFS+ILEN)
    ENDIF
    ILENS = D%NLEN_M(JW,2,JM)*IFIELD
    IOFFS = (D%NSTAGT0B(JW)+D%NOFF_M(JW,2,JM))*IFIELD
    IOFFR = (D%NSTAGT0BW(JW,MYSETW)+D%NOFF_M(JW,2,JM))*IFIELD
    FOUBUF_C(IOFFR+1:IOFFR+ILENS)[IPE]= &
           & FOUBUF_IN(IOFFS+1:IOFFS+ILENS)
  ENDDO
ENDDO
!$OMP END PARALLEL DO
SYNC IMAGES(D%NMYSETW)
FOUBUF(1:IBLEN)=FOUBUF_C(1:IBLEN)[MYPROC]
```

Semi-Lagrangian advection

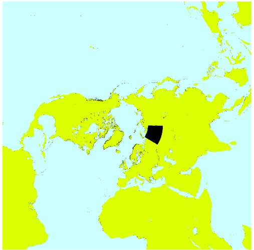




- Computes the trajectory of each grid-point backward in time
- Interpolates advected quantities at departure and mid-point
- Weather dependent communication pattern

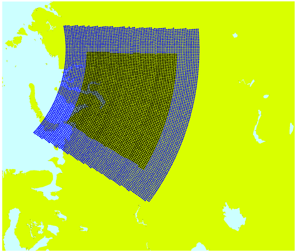


Semi-Lagrangian advection



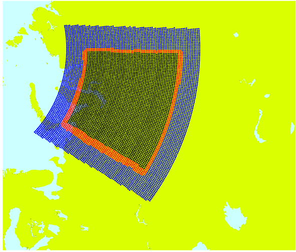
- Task 11 encountered highest wind-speed of 120 m/s (268 mph) during 10 day forecast starting 15 Oct 2004

Semi-Lagrangian advection



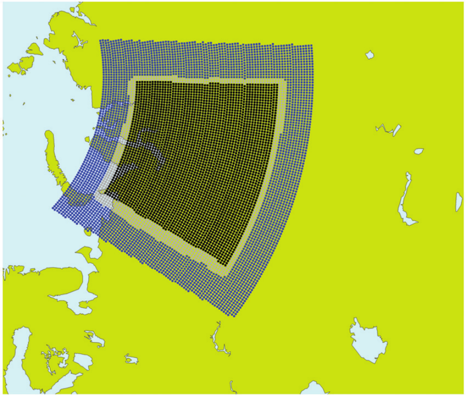
- Halo width assumes a maximum wind speed of 400 m/s x 720s (timestep size) = 288km
- Get u,v,w wind vector variables (3) from ‘neighbour’ tasks to determine departure and mid-point of trajectory

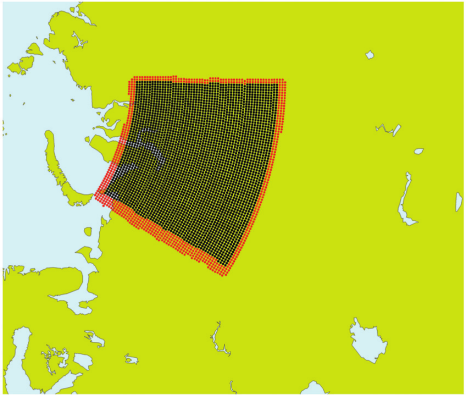
Semi-Lagrangian advection



- Get rest of the variables (26) from the red halo area and perform interpolations
- Note that volume of halo data communicated is dependent on wind speed and direction in locality of each task

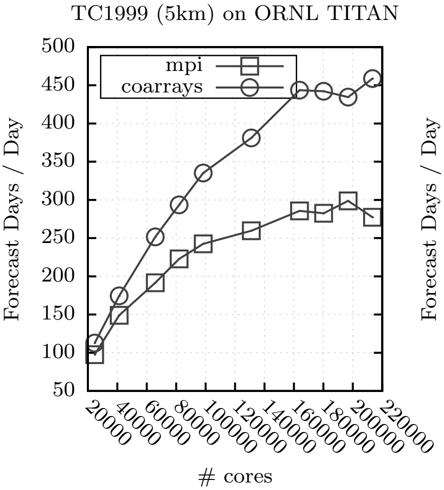
One-sided comms in spectral transforms



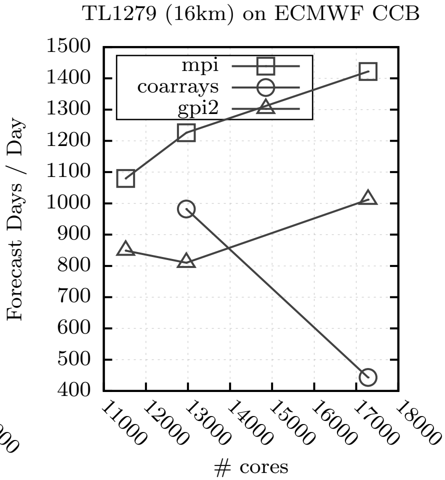


- Only get grid-point columns as and when needed from neighbouring tasks therefore removing the need for very large halos
- Single OpenMP loop for computing departure point and interpolations

Results



TC1999 (5km) on ORNL TITAN
Forecast Days / Day
cores
mpi
coarrays
500
450
400
350
300
250
200
150
100
50
20000
40000
60000
80000
100000
120000
140000
160000
180000
200000
220000



TL1279 (16km) on ECMWF CCB
Forecast Days / Day
cores
mpi
coarrays
gpi2
1500
1400
1300
1200
1100
1000
900
800
700
600
500
400
11000
12000
13000
14000
15000
16000
17000
18000

Results

- Fortran2008 Coarray SYNC had some serious performance issues if more than one electrical group (384 nodes) was used on our Cray XC40
- Cray DMAPP library was not thread safe as at the time in CCE 8.0.6
- This was not a problem on ORNL Titan though (Cray XK7 with Gemini interconnect)
- Interestingly, GPI2 did not see the same issues even though it used identical comp-comm overlap scheme
- Lots of compiler bugs found when using Fortran08 Coarrays
- 25% speed-up on ORNL Titan vs MPI implementation at 220K cores.

Mozdzynski, M. Hamrud, N. Wedi, J. Doleschal, and H. Richardson. A PGAS implementation by co-design of the ECMWF Integrated Forecasting System (IFS). In 2012 SC Companion: High Performance Computing, Networking Storage and Analysis, pages 652–661. IEEE, 2012.

Mozdzynski, M. Hamrud, N. Wedi. A Partitioned Global Address Space implementation of the European Centre for Medium Range Weather Forecasts Integrated Forecasting System. The International Journal of High Performance Computing Applications 2015, Vol. 29(3) 261–273.

Conclusions

- First attempt at one-sided communication in IFS led to promising results
- At the time, main issue revolved around the available system stack
- We are keen to have another go this time using MPI RMA as well as OpenSHMEM, UCX etc
- We need to go further than simply extending parallelism within threaded regions to changing algorithms to exhibit more potential for overlapping computation and communication
- For example, rewrite the Semi-Lagrangian scheme to take advantage of task parallelism
- To help the above, features such as notifications as found in GASPI/GPI2 are very useful